自然言語処理 128－4
（1998. 11. 5）

学習による文節まとめあげ
— 決定木学習，最大エントロピー法，用例ベースによる手法と排反な規則を用いる新手法の比較 —

村田 真樹　内元 清貴　馬 青　井佐原 均

郵政省 通信総合研究所

〒651-2401 神戸市西区岩岡町岩岡 588-2

tel:078-969-2181 fax:078-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,uchimoto,qma,isahara}@crl.go.jp

あらまし

言語コーパスの増加により教師あり学習の研究が極めて重要になってきている．本研究では普遍的に最もよい教師あり学習のアルゴリズムの作成に向け，文節まとめあげという簡単な問題を対象として，既存の決定木学習，最大エントロピー法，用例ベースの手法と排反な規則を用いる新手法の比較実験を行なった．その結果，今回の問題設定では排反な規則を用いる新手法が最もよいことがわかった．

キーワード　文節，機械学習，決定木，最大エントロピー法，用例ベースのアプローチ

Machine Learning Approach to Bunsetsu Identification
— Comparison of Decision Tree, Maximum Entropy Model, Example-Based Approach, and A New Method Using Category-Exclusive Rules —

Masaki Murata　Kiyotaka Uchimoto　Qing Ma　Hitoshi Isahara

Communications Research Laboratory,

Ministry of Posts and Telecommunications

588-2, Iwaoka, Nishi-ku, Kobe, 651-2401, Japan

tel:+81-78-969-2181 fax:+81-78-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,uchimoto,qma,isahara}@crl.go.jp

Abstract

Research into supervised learning is extremely important because of the increased linguistic corpora. In order to achieve the best supervised learning, we carried out experiments on bunsetsu identification by comparing the three existing methods (decision tree, maximum entropy model, and example-based approach) and our new method using category-exclusive rules. In these experiments our new method using category-exclusive rules was better than the other learning methods.

key words　Bunsetsu, Machine Learning, Decision Tree, Maximum Entropy Model, Example-Based Approach

1 はじめに

近年，言語関係のコーパスが増加するとともにコーパスベースのアプローチによる言語処理の研究が盛んになってきている．また，さまざまな学習アルゴリズムが言語処理の問題に対しても用いられるようになってきており，正解タグつきコーパスを用いた機械学習に関する研究も異常なほどに盛んになってきている．

機械学習の方がよいか，人手によるルールベースのアプローチの方がよいかと常々議論されてきてはいるが，それぞれの利点と欠点をまとめると，以下のようになると思われる．

	機械学習	人手ルールベース
利点	メインテナンスが楽 容易に類似問題に適用可能	人間の知識や経験を利用したり微細に調節できたりするため大抵は精度がよい はっきりした言語知識が得られる場合がある
欠点	精度が良ければいうことはないが，えてして精度が悪い 学習結果を人間の理解容易なものとして得ることが難しい	メインテナンスが大変 他の問題への適用困難

言語コーパスが増加する現在，コーパスが増加すると学習データが増え機械学習の解析精度はどんどん向上すると予想されるので，コーパスを用いた機械学習の研究はますます重要になってくると思われる．しかし，本当に機械学習の方が人手によるルールベースよりもよいのか．また，機械学習の手法も多種多様であり，どの機械学習の手法が最も有力かなど，根本的なところでさえ明らかでない状況である．このような状況でやみくもに様々な問題に対して好き勝手な機械学習の手法で研究を行なうことは問題を複雑にするだけでなく，ものごとの本質を見失わせることになるであろう．

本稿ではこのような状況を少しでも改善させるために，正解タグ付きコーパスが存在し，なおかつ，自然言語処理の問題として最も簡単な問題のひとつであろう文節まとめあげの問題を研究の対象とした．研究の対象として簡単な問題を選んだ理由は，簡単な問題であれば，複数の手法の有用性を調べることが容易であろうと考えたためである．また，教師あり機械学習の問題は，それ自体正解データを利用できるという類似性で一つにまとめて考えることができ，一つの教師あり機械学習の問題が解ければ他の教師あり機械学習の問題もある程度容易に解けるであろうと推測されるからである．このようなときにわざわざ難しい問題をとりあげるのは，全く非効率的であるとしか思われない．本研究は，文節まとめあげの問題という簡単な問題を足掛かりに，いろいろな問題を統一的に扱えるような，教師あり機械学習の手法というものを模索する試みの出発点となっている．

本研究で強調しておきたいことをあらかじめ整理しておくと以下のようになる[1]．

- 文節まとめあげという問題自体を機械学習の問題として初めて取り扱った研究である[2]．
- 既存の有力な手法 (決定木学習，最大エントロピー法，用例ベースのアプローチ (類似度の利用)) を用いて実験を行ない，本研究の問題設定では，それぞれの手法の優劣はおおよそ以下のようだと言えそうな結果を得た．

  用例ベース $\geq$ 最大エントロピー $\geq$ 決定木学習
- 後節で述べる排反な規則を用いる新手法を提案し，用例ベースの手法と同程度かもしくは若干高い精度をあげた．

2 文節まとめあげの問題

本研究の立場は，文節まとめあげという簡単な問題で種々の教師あり学習の実験を行ない以下の各手法の傾向，利点，弱点をみきわめることにある．

- 決定木学習
- 最大エントロピー法
- 用例ベース (最も類似した用例の利用)
- 新手法 1(排反な規則の利用)
- 新手法 2(排反な規則と類似度の利用) [3]

文節まとめあげの処理は，形態素解析と構文解析の間の処理である．本稿では形態素解析まではなされたものとし，形態素解析の結果から形態素の情報を取り出しそれを用いて文節の認定 (文節まとめあげ) を行なう．本稿では文節まとめあげの問題を以下のように形態素間に文節を区切るための記号 "|" を挿入する問題として扱う．

(例) 文節を | まとめあげる

処理の手順は，文の頭から各形態素のすき間に対してまわりの形態素の情報によって文節を区切るための記号 "|" を挿入するかいなかを判定していくということになる (本研究では簡単のため，挿入した区切り記号などの情報は用いない)．

[1] 筆者は最近論文においてはこのような研究の重要な点を序論において2～4個の箇条書としてまとめるべきではないかと考えている．このような項目が存在していると，その研究の新規性や価値が読者にとって容易に理解することができ便利であると考えている．

[2] 森の研究(1) は構文解析の際に同時に文節まとめあげを確率モデルで解析しているが，文節まとめあげという問題自体を扱っている研究ではない．また，文節まとめあげの精度評価を行なっていない．

[3] 「あらまし」「はじめに」「おわりに」で述べている「排反な規則を用いる新手法」はこの新手法2を意味することに注意．

	文	を	区切る	.
品詞	名詞	助詞	動詞	特殊
品詞細分類	普通名詞	格助詞	基本形	句点
意味情報	×	none	217	×
単語自体	×	を	区切る	×

図 1: 解析に用いる情報

文節を区切るかいなかの判定に用いる情報は，文節を区切るための記号 "|" を挿入するかいなかを判定する形態素のすき間の前二形態素と後ろ二形態素の形態素情報である．形態素情報として用いるものは，各形態素ごとに下記の四つのものである．

1. 品詞[4]
2. 品詞細分類 (or 活用形) [5]
3. 意味情報 (分類語彙表(3) の分類番号上位 3 桁)
4. 単語自体 (語尾変化する単語の場合は基本形を用いる)

ただし，解析に用いる情報が増え過ぎると解析に時間がかかり実験の不都合が生じるため，本稿では両端二形態素については，意味情報, 単語自体の情報は用いない．

図 1は，「文を区切る．」という文の「を」と「区切る」という形態素のすき間に文節区切り記号を挿入するかいなかを判定する際に用いる情報を示している．「文」「を」「区切る」「．」に対してそれぞれ品詞の情報として「名詞」「助詞」「動詞」「特殊」がある．また，品詞細分類として同様に「普通名詞」「格助詞」「基本形」「句点」の情報がある．また，意味情報，単語自体は上で述べたように実験の都合上真中よりの二つの形態素「を」「区切る」に対してのみ用いる．意味情報としては分類語彙表の分類番号上位 3 桁を用い，「を」は分類語彙表になく，ないという意味の印 "none" の情報を，「区切る」は分類語彙表にあり上位三桁の "217" の情報を用いる．単語自体はそのまま，「を」「区切る」となる．

こういった情報を用いてそれぞれの形態素のすき間に対して文節区切り記号を挿入するかいなかを判定していくことで解析を行なう．

3 各種学習手法ごとの文節まとめあげへの適用法

3.1 決定木学習

決定木学習とは簡単にいうと分類決定のための yes/no の分岐の木を学習するアルゴリズムのことである．本研究の決定木学習には，Quinlan の C4.5(4) を用いる．

[4] 本研究では品詞体系は juman3.5(2) に準拠している．

[5] 本研究では情報の圧縮のために品詞細分類の欄には，活用する単語については活用形を記述している．しかしこのようにすると品詞細分類を書く場所がなくなってしまうので，活用形がある単語で品詞細分類の項目もある単語については，品詞細分類の情報は品詞の情報と合わせて品詞の欄に記載することにしている．

解析に用いる情報としては，前節で述べたとおり品詞と品詞細分類と意味情報と単語自体の四つである．これをそれぞれ決定木学習の際の属性として用いる．ただし，両端二形態素については前節で述べた通り品詞と品詞細分類しか用いないので，学習の際に用いられる属性の数は下記のように 12 個となる．

$$\text{属性数} = 2 + 4 + 4 + 2 = 12$$

二つ前の形態素	一つ前の形態素	一つ後の形態素	二つ後の形態素
品詞, 細分類	品詞, 細分類, 意味情報, 単語自体	品詞, 細分類, 意味情報, 単語自体	品詞, 細分類
2個	4個	4個	2個

例えば，問題となっている部分が図 1のようになっている場合は，"属性「二つ前の形態素の品詞」の属性値は名詞" といった情報が 12 個用いられることになる．

3.2 最大エントロピー法 (ME)

最大エントロピー法はデータスパースネスに強い方法で，最近になって多くの研究者によって用いられるようになってきている(5)(6)(7)(8)．本研究の最大エントロピー法の実験では，文献(9) のシステムを用いる[6]．解析は，そのシステムの出力から区切りの確率と区切らない確率を計算し，その確率の大きい方であると推定することによって行なう．

最大エントロピー法でも決定木学習と同じく解析に用いる情報は，品詞と品詞細分類と意味情報と単語自体の四つである．ただし，決定木学習とは異なり最大エントロピー法は素性 (情報) の AND(組み合わせ) の情報をシステムが自動的には考えないので，素性として与えるときにあらかじめ情報を組み合わせておく必要がある．

まず，各形態素内の情報，すなわち，品詞と品詞細分類と意味情報と単語自体の組合せを考える．四つの情報があるので，あらゆる組合せを考える場合 $2^4 - 1$ 通りの組合せを考えることになる．しかしこれでは，組合せの数が多く計算機の負荷が大きくなるので，品詞と品詞細分類と意味情報と単語自体の情報がこの順に細かくなる性質を利用して，本研究では四つの情報の組合せとしては以下の 4 種類を用いることにした．

情報 A：品詞
情報 B：品詞と品詞細分類
情報 C：品詞と品詞細分類と意味情報
情報 D：品詞と品詞細分類と意味情報と単語自体 (1)

ただし，両端二形態素については前節で述べた通り品詞と品詞細分類しか用いないので，上記の 4 種類のうちの

[6] 今は Web 上に存在していない．

上の二つのみとなる.

次に各形態素間の組合せを考えると，以下のように情報の組合せの数は $2 \times 4 \times 4 \times 2 = 64$ となる.

二つ前の形態素　一つ前の形態素　一つ後の形態素　二つ後の形態素

$$\left\{\begin{array}{l}\text{情報 A}\\ \text{情報 B}\end{array}\right\} \& \left\{\begin{array}{l}\text{情報 A}\\ \text{情報 B}\\ \text{情報 C}\\ \text{情報 D}\end{array}\right\} \& \left\{\begin{array}{l}\text{情報 A}\\ \text{情報 B}\\ \text{情報 C}\\ \text{情報 D}\end{array}\right\} \& \left\{\begin{array}{l}\text{情報 A}\\ \text{情報 B}\end{array}\right\}$$

2個　4個　4個　2個

これらの組合せの他にデータスパースネスの対策として，両端二形態素の情報がそれぞれ片方しか用いられない場合，両方用いられない場合，真中二形態素のうち片方が用いられない場合を考えると，最大エントロピー法で用いられる素性の数は一箇所につき下記のように 152 個となる[7].

$$\begin{aligned}\text{素性の数} = {} & 2 \times 4 \times 4 \times 2\\ & + 2 \times 4 \times 4\\ & + 4 \times 4 \times 2\\ & + 4 \times 4\\ & + 4\\ & + 4\\ & = 152\end{aligned}$$

例えば，問題となっている部分が図 1のようになっている場合は，四つの形態素の情報として二つ前の形態素の情報 B と，一つ前の形態素の情報 D と，一つ後の形態素の情報 C と，二つ後の形態素の情報 A を用いるような素性は，“名詞: 普通名詞 | 助詞: 格助詞:none: を | 動詞: 基本形:217| 特殊: 句点” といったものとなり，このような素性を 152 個用いることになる.

3.3 用例ベース (類似度の利用)

用例ベースのアプローチとは， Nagao [10] によって 1984 年に機械翻訳の問題において初めて提案された手法で，ある解析を行なう際にそれと最も類似した用例を持ってきてその用例を利用して解を得るという手法である. 近年になって文献[11][12] において格フレーム選択や照応解析など機械翻訳以外の問題にも用いられるようになってきた. また，文献[13] も形態素解析の問題を用例ベースのアプローチで解いている研究といえよう. 本節ではこの用例ベースのアプローチを利用して文節区切りを行なう手法を述べる.

この用例ベースの手法でも公平なように用いる情報は最大エントロピー法と同じく各形態素の情報としては，式 (1) にあげた情報 A，情報 B，情報 C，情報 D の 4 種類を用いる.

[7] この手法で例えば京大コーパス 95 年 1 月 1 日分 (形態素間のすき間の数が 25,814 のもの) で素性を取り出すとその種類の数は 1,534,701 個となる.

		文	を	区切る	.
	s(x)	m_{-2}	m_{-1}	m_{+1}	m_{+2}
情報なし	1	—	—	—	—
情報 A	2	名詞	助詞	動詞	特殊
情報 B	3	普通名詞	格助詞	基本形	句点
情報 C	4	×	none	217	×
情報 D	5	×	を	区切る	×

図 2: 類似度の説明のための例

また，用例ベースでは入力と用例との類似度を定義してやる必要がある. なるべく微妙な違いまで検出できるような類似度を設定するために，入力と用例の類似するレベルとして，データスパースネスのことを考慮した最大エントロピー法と同じ 152 パターンを考えることにする. 本稿では入力と用例の類似度 S を，それらの一致するレベルがこの 152 のパターンのどのレベルになっているかに応じて以下のように定める.

$$\begin{aligned}S = {} & s(m_{-1}) \times s(m_{+1}) \times 10000\\ & + s(m_{-2}) \times s(m_{+2})\end{aligned} \quad (2)$$

ただし，m_{-1}，m_{+1}，m_{-2}，m_{+2} は，解析している形態素間のすき間に対して，一つ前の形態素，一つ後の形態素，二つ前の形態素，二つ後の形態素を意味する. また，s(x) は，形態素 x の形態素単位の類似度で以下のように定義される.

$$\begin{aligned}s(x) = {} & 1(\text{形態素 x の情報が全く一致しない場合})\\ & 2(\text{形態素 x の情報 A のレベルでのみ一致する場合})\\ & 3(\text{形態素 x の情報 B のレベルで一致する場合})\\ & 4(\text{形態素 x の情報 C のレベルで一致する場合})\\ & 5(\text{形態素 x の情報 D のレベルで一致する場合})\end{aligned} \quad (3)$$

式 (2) は真中の二形態素が特に重要であると考えて作成したものである. 本稿ではこの類似度の式を用いるが，もっとよい類似度の式があるかもしれないし，また，類似度自体なんらかの学習アルゴリズムで定めるということをした方がいいかもしれない[8].

類似度の例を示しておく. 例えば，問題となっている部分が図 2のようになっている場合，152 パターンのうち，“名詞 | 助詞 | 動詞 | 特殊” といった品詞レベルですべて一致し，より細かいレベルでは一致しないパターンの場合は $2 \times 2 \times 10000 + 2 \times 2$ で類似度は 40004 と

[8] 意味的な要素を考慮して類似用例を取ってくる際に決定木学習を用いている研究がある. Jiri[14] の研究では前置詞句のかかり受けの問題に矢田[15] の研究では「A の B」の意味解析に利用している. それぞれの研究とも，アドホックな類似式を作成せず，例えば名詞 A の意味がこういう場合は名詞 B よりも名詞 A の類似性を優先するなどといったことが可能なように単語の意味シソーラス上で決定木学習を行なう興味深い研究である.

なる．また，“名詞|助詞:格助詞:を|—|—”といったパターンで一致する場合は5 × 1 × 10000 + 2 × 1で類似度は50002となる．

本手法はこの類似度の値が最も高い用例を取り出してその用例が文節区切りになっていれば，文節区切り記号を挿入すると判定し，そうでない場合は文節区切り記号を挿入しないと判定する．類似度が等しい用例が複数あった場合，どの用例を利用して解析すればよいかが曖昧な場合がある．本稿の実験では，類似度が等しい用例が複数あった場合，その用例の集合において文節区切りになった数とそうでなかった数を調べ，多かった方であると推定するようにしている．

3.4 新手法1(排反な規則を利用)

前節までは，文節まとめあげに用いる手法として，有力な既存の三つの手法を説明してきた．本節と次節では，われわれが新たに考えた手法について説明する．

以上までの節の議論からすると，解析に用いる情報は今の条件で最大限有効に利用するには，先の二手法と同じように形態素のすき間1箇所につき152種類のパターンを考えるのが良さそうである．そこで，学習セットのデータのすべてのすき間において，152のパターンのデータの統計情報をとってみる．これにより，例えば，“名詞:普通名詞|助詞:格助詞:none:を|動詞:基本形:217|特殊:句点”というパターンが学習セット中に13回出現し，そのうち10回は文節の区切り記号を挿入すべき場所であるといったデータが得られる．ここでは，以下のようなパターンの情報が得られたとしよう．下記のデータ中の区切部分は文節の区切り記号を挿入すべき場所であることを意味し，データ中の継続部分は文節の区切り記号を挿入すべきでない場所であることを意味する．

パターンA	区切部分	10回	継続部分	3回
パターンB	区切部分	100回	継続部分	23回
パターンC	区切部分	33回	継続部分	0回
パターンD	区切部分	230回	継続部分	310回
.....	.....		.....	

この統計情報において，パターンCは33回中33回とも，区切部分となっていたということを示している．

このパターン情報は確率で表現すると以下のような確率という確信度を持った規則のようなものとなる．

規則A	区切部分になる確率	76.9% (10/ 13)
規則B	区切部分になる確率	81.3% (100/123)
規則C	区切部分になる確率	100% (33/ 33)
規則D	継続部分になる確率	57.4% (310/540)
.....	.....	.....

例えば，規則Cは確率100%で区切部分と推定する規則を意味し，規則Dは確率57.4%で継続部分と推定する規則を意味するようになっている．

このような規則が並んでいる際，どの規則を信用するかをまかされた場合は確率が最も大きい規則Cを信用して区切部分と推定したくなるだろう．本節の新手法1(排反な規則の利用)では，この方法をとるものである．すなわち，最も確率の高い規則(パターン)にしたがって解析するものである．ここで，確率が100%になっている規則を，区切か継続かのいずれかのみという排他的な集合に分割されている場合の規則という意味で，排反な規則と呼ぶこととする．本問題では適用される規則は1箇所につき最大で152個もあるので，排反な規則が適用されて推定されることが多く，本節の新手法1は，排反な規則を利用する手法とも呼ぶこととする．

確率の等しい規則が複数あった場合，どの規則を利用して解析すればよいかが曖昧な場合がある．本節の手法では確率が等しい用例が複数あった場合は，その規則の統計データでの誤り数(例えば，規則Aなら3)の小さいものを選ぶ．その値も等しい場合はその同点であった規則の集合において，継続を支持する規則の数と区切を支持する規則の数を数えあげ，その数の多い方であると推定するようにしている．また，排反な規則(確率が100%の規則)が複数適用される場合は，出現回数が頻度1の規則は信用できないとし省いて計算している．このあたりの処理はあまり深く検討しておらず改良の余地が残っている．

3.5 新手法2(排反な規則と類似度を利用)

本節の手法は前の二節の手法，すなわち，類似度を用いる手法と，排反な規則を用いる手法を組み合わせた方法である．

解析に用いる情報はいままでと同じ152パターンである．この152パターンを前節と同様に規則として扱い，最も確率の高い規則を信用して解析する．最も確率の高い規則が複数ある場合は，用例ベースの手法の節の類似度の値を用いこの値が最も高い規則を信用して解析する．また，その値が等しい規則が複数ある場合は新手法1と同様にその規則の統計データでの誤り数などに応じて信用する規則を選択する．

4 実験および考察

実験は京大コーパス[16]の毎日新聞95年1月1日～95年1月5日の記事で行なった．システムの入力となる形態素の情報はコーパスに付与されているものを用いた．

まずどの教師あり学習の手法が最も有効かどうかを調べるために，以下の二つの実験を行なった．

- 実験1

  95年1月1日を学習セット，95年1月3日をテストセットとする実験
- 実験2

表 1: 実験 1 での学習セット (1 月 1 日) での解析精度

手法	F	再現率	適合率
決定木	99.58%	99.66%	99.51%
ME	99.20%	99.35%	99.06%
用例ベース	99.98%	100.00%	99.97%
新手法 1	99.98%	100.00%	99.97%
新手法 2	99.98%	100.00%	99.97%
knp 2.0b4	99.23%	99.78%	98.69%
knp 2.0b6	99.73%	99.77%	99.69%

形態素間のすき間の数 25,814. 区切部分の数 9,523.

表 2: 実験 1 でのテストセット (1 月 3 日) での解析精度

手法	F	再現率	適合率
決定木	98.87%	98.67%	99.08%
ME	98.90%	98.75%	99.06%
用例ベース	99.02%	98.69%	99.36%
新手法 1	98.98%	98.49%	99.48%
新手法 2	99.12%	98.82%	99.43%
knp 2.0b4	99.13%	99.72%	98.54%
knp 2.0b6	99.66%	99.68%	99.64%

形態素間のすき間の数 16,983. 区切部分の数 6,166.

表 3: 実験 2 での学習セット (1 月 4 日) での解析精度

手法	F	再現率	適合率
決定木	99.70%	99.71%	99.69%
ME	99.07%	99.23%	98.92%
用例ベース	99.99%	100.00%	99.98%
新手法 1	99.99%	100.00%	99.98%
新手法 2	99.99%	100.00%	99.98%
knp 2.0b4	98.94%	99.50%	98.39%
knp 2.0b6	99.47%	99.47%	99.48%

形態素間のすき間の数 27,665. 区切部分の数 10,143.

表 4: 実験 2 でのテストセット (1 月 5 日) での解析精度

手法	F	再現率	適合率
決定木	98.50%	98.51%	98.49%
ME	98.57%	98.55%	98.59%
用例ベース	98.79%	98.56%	99.02%
新手法 1	98.74%	98.27%	99.22%
新手法 2	98.88%	98.62%	99.14%
knp 2.0b4	99.07%	99.43%	98.71%
knp 2.0b6	99.51%	99.40%	99.61%

形態素間のすき間の数 32,304. 区切部分の数 11,756.

95 年 1 月 4 日を学習セット, 95 年 1 月 5 日をテストセットとする実験

3.4節,3.5節で述べた新手法は上記の実験 1 を試行して作成したので, 実験 1 は若干クローズデータの意味合いがあるため, 実験 2 を行なっている.

実験結果を表 1〜表 4に示す. 表では, 3節で述べた 5 つの手法の他に比較のために knp 2.0b4[17] と knp 2.0b6[18] の精度も示しておいた. knp の結果については表に示す「学習セット」「テストセット」に意味はない. knp で実験する際も knp の入力となる形態素の情報はコーパスから得ている. 95 年 1 月 5 日の実験では knp の出力が不完全な文があったが, この文はすべての手法の実験で除いた. 表中の "F" は F measure を意味し, 再現率, 適合率の逆数を足して二で割ったものを逆数にした値のことである. 再現率, 適合率は区切部分に対するもので, それぞれシステムの正解した区切部分の個数をすべての区切部分の個数で割ったもの, システムの正解した区切部分の個数をシステムが区切部分と推定したものの個数で割ったものを意味する. 最大エントロピー法 (ME) での実験では文献[9] のシステムを用いたが, 低頻度の素性が存在している場合システムが動かなかったので, "ある素性" と "その素性が存在するときに区切るかいなかの情報" の組み合わせ[9]の頻度が 11 以下のものは消して実行した (頻度が 10 以下のものを消して実行した場合システムが動かなかったので, しかたなくこのようにした). このため, これらの素性を削除しない場合最大エントロピー法による方法ではここに示した精度以上のものが出る可能性がある. また, このシステムではパラメータの繰り返し学習が行なわれるが, 本研究では繰り返し回数はとりあえず 200 回に固定して実験を行なった. また, 決定木学習においては, C4.5 のシステムを用いたが実験においては -s オプションをつけて実行した. -s オプションをつけると属性値のグループ化が行なわれるが, 属性値の種類の数が大きいと計算にかなりの時間がかかる. そこで, 頻度が 10 未満の属性値については「その他」という属性値を用意しすべてその属性値を割り当てて実験を行なった.

表 1〜表 4の実験結果から以下のようなことがわかる.

- テストセットにおいては, 決定木よりも最大エントロ

[9] 最大エントロピー法のプログラムの入力として, "素性" と "その素性が存在するときに区切るかいなかの情報" の組みあわせの頻度が必要となる. この頻度が小さい組み合わせを省くとシステムが動作するようになったので, 本研究ではそのようにして実験を行なった. しかし, 動作させるためにはもっとよい方法があるかもしれない.
また, 素性と区切るかいなかの情報の組み合わせで捨てているので, 単純に素性の頻度で捨てる場合に比べて, より多くの素性を捨てていることになる.

表 5: 追加実験 1：用いる意味情報の変更

分類番号の利用桁数	F	再現率	適合率
3桁	99.12%	98.82%	99.43%
5桁	99.14%	98.88%	99.41%
7桁	99.11%	98.86%	99.36%

表 6: 追加実験 2：両端二形態素の単語・意味情報の利用

両端の単語・意味情報	F	再現率	適合率
利用せず	99.14%	98.88%	99.41%
利用	99.17%	98.91%	99.43%

ビー法の方が若干ではあるが精度がよかった[10]. 最大エントロピー法の欠点として，自動的に素性の組合せを学習しないというものがあるが，本実験での最大エントロピー法ではあらゆる素性の組合せを用いていたので，その欠点が克服され決定木よりも精度が良くなったと思われる.

- 現時点では，用例ベースの方が最大エントロピー法よりも精度が高かった. ただし，最大エントロピー法はシステムの都合で素性を削除しないとシステムが動作しなかっただけなので，これらの素性も利用できるようになると精度が向上する可能性がある.
- 新手法 1(単純に排反な規則を用いる手法) は用例ベースよりも精度が低かったが，新手法 2(排反な規則と類似度を用いる手法) は学習アルゴリズムの中では最も高い精度を得た.
- 用例ベース，新手法 1，新手法 2 の三つの手法は学習セットにおいては 100% に極めて近い精度を出している. これらの手法は特に学習セットにおいては強いことがわかる.
- knp でも解析してみたが，テストセットにおける精度は knp のものが最も高かった.
- knp においては，knp 2.0b4 と knp 2.0b6 の二つのバージョンのもので実験を行なったが，これは knp 2.0b6 の方が断然精度が良かった. 人手によるシステム向上もかなり効果的であることがわかる.

以上の実験から今のところ新手法 2 がいろいろな条件つきではあるが，機械学習の手法としては最も優れていそうであると思われる.

次に解析に用いる情報を増やすことで新手法 2 の精度が向上しないかどうかを試すために以下の三つの追加実験を行なった.

追加実験 1：用いる意味情報の変更

今までの実験では意味情報としては分類語彙表の分類番号の 3 桁を用いていたが，より多くの桁を用いた場合精度がどのようにかわるのかを調べた. 95 年 1 月 1 日を学習セット，95 年 1 月 3 日をテストセットとして，意味情報を分類語彙表の分類番号の 3 桁，5 桁，7 桁と変化させて実験を行なった. その結果を表 5に示す. 表のように精度はあまり大差はないが，分類語彙表の 5 桁が最も精度が良かった. 以降の実験では分類語彙表の 5 桁を意味情報として用いる.

追加実験 2：両端二形態素の単語情報，意味情報の利用

今までの実験ではシステムの負荷の軽減のために両端二形態素については単語情報，意味情報を用いていなかったが，それも用い解析に用いる情報を増やすことで精度向上を目指す. 95 年 1 月 1 日を学習セット，95 年 1 月 3 日をテストセットとして，両端二形態素の単語情報，意味情報も利用して実験を行なった. その結果を表 6に示す. 表のように精度の向上は F measure で 0.03% であり，両端二形態素の単語情報，意味情報の追加の効果が小さいことがわかる. しかし，それでも用いないよりは用いた方が精度が高くなることがわかった.

追加実験 3：学習データの量

追加実験 1,2 より，意味情報としては分類語彙表の 5 桁を用い，両端二形態素の単語情報，意味情報についても加えて用いるのが良いことがわかったので，本実験ではそのような条件で実験を行なうことにした. その条件で，学習セットのデータ量を変化させて精度がどのように変わるかを調べる実験を行なった. 計算機の負荷を考慮し，ここでは学習セットを減少させる実験を行なった. テストセットを 95 年 1 月 3 日に固定して学習セットを 95 年 1 月 1 日のものから，それの約 1/2，約 1/4，約 1/8 のもので実験を行なった. その結果を表 3と図 3に示す. 学習データが増加すると精度が向上することがわかる. まだコーパスを増やしても少しは精度があがりそうな状態である.

本研究では用例ベースや排反な規則を用いる手法が良さそうな結果を得た. しかし，問題が難しくなり考慮すべき情報がたくさんある問題では，あらゆる情報の組合せを扱うことが難しくなり，用例ベースや排反な規則の方法を用いることが困難になってくる. また，12 個の属性しか用いていない決定木の手法では，まだまだ属性を増やして実験をすることが可能であり，解析に用いる情

[10] 本実験ではともに素性を捨てているので正確な比較になっていない. しかし，決定木では形態素ごとの情報で捨てていて，最大エントロピー法は 4 形態素の AND を取った後の情報 (厳密にはさらに区切るかいなかの情報との組み合わせをとった情報) で捨てており，同じ頻度で捨てる場合であっても最大エントロピー法の方がより多くの素性を捨てることになっている. このため，若干でも精度がいい最大エントロピー法の方が決定木よりもよさそうであると推測される.

表 7: 追加実験 3：学習データの量の変化

学習データの量 (すき間の数)	F	再現率	適合率
1/8 のデータ (3,289 個)	98.07%	97.86%	98.29%
1/4 のデータ (6,564 個)	98.56%	98.26%	98.87%
1/2 のデータ (13,053 個)	98.92%	98.70%	99.14%
1 日分すべて (25,814 個)	99.17%	98.91%	99.43%

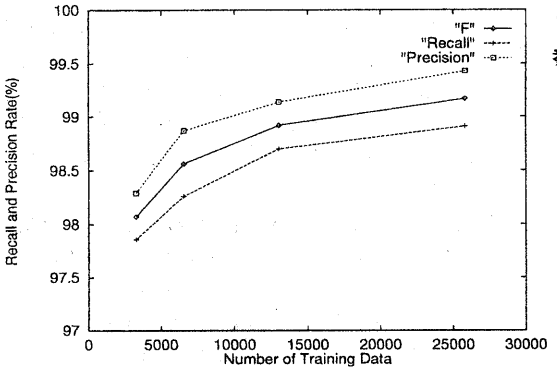

図 3: 追加実験 3：学習データの量の変化

表 8: knp で誤ったが新手法 2 で正解した例

コツコツ|不足我慢し、
余裕を|持って|不足退けた
会社を|グループ分け|×して|、
最も|慣れ|×親しんでいる

報を増やすことで精度向上を行なえる可能性があるが，用例ベースや排反な規則を用いる手法では現在の計算機のパワーでは難しい．それでもこれらのことは計算機技術の向上により克服されるであろうから，解析に用いる情報が等しい場合に精度がどのようになるかを調べる本研究のようなアプローチは重要である．

最後に，knp で誤ったが新手法 2 では正解した例を表 8に示しておく．表中「不足」印がついている部分は knp が誤って区切らなかったものを意味し，×印がついている部分は knp が誤って区切ったものを意味する．実験 1 のテストセットでは knp2.0b6 の F measure は 99.66% であったが，knp2.0b6 と新手法 2 のどちらかが正解していれば正解とするとき F measure は 99.83% となった．knp2.0b6 は精度がよいが，新手法 2 で正解して knp2.0b6 で誤るものも少しはあることがわかる．

5 おわりに

本研究では，文節まとめあげという比較的簡単な問題で，種々の教師あり機械学習の比較を行なった．本研究の実験の結果では，既存の手法の間には下記のような優劣があった．

用例ベース ≥ 最大エントロピー ≥ 決定木学習

また，排反な規則を用いるという新しい手法を提案し，それが用例ベースの手法と同程度かもしくは若干高い精度をあげた．

われわれは，今後この排反な規則を用いる手法を発展させて，あらゆる教師あり学習の問題を高精度で解くような手法の開発を行ないたいと思っている．

参考文献

(1) Shinsuke Mori and Makoto Nagao. A Stochastic Language Model using Dependency and Its Improvement by Word Clustering. *COLING '98*, pp. 898–904, 1998.

(2) 黒橋禎夫, 長尾眞. 日本語形態素解析システム JUMAN 使用説明書 version 3.5. 京都大学大学院工学研究科, 1997.

(3) 国立国語研究所. 分類語彙表. 秀英出版, 1964.

(4) J.R. キンラン. AI によるデータ解析. トッパン, 1995.

(5) Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71, 1996.

(6) Adwait Ratnaparkhi. A Maximum Entropy Model for Part-Of-Speech Tagging. *Proceedings of Empirical Method for Natural Language Processings*, pp. 133–142, 1996.

(7) Adwait Ratnaparkhi. A Linear Observed Time Statistical Parser Based on Maximum Entropy Models. *Proceedings of Empirical Method for Natural Language Processings*, 1997.

(8) Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 152–160, 1998.

(9) Eric Sven Ristad. Maximum Entropy Modeling Toolkit, Release 1.6 beta. http://www.mnemonic.com/software/memt, 1998.

(10) Makoto Nagao. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. Artificial and Human Intelligence, pp. 173–180, 1984.

(11) Sadao Kurohashi and Makoto Nagao. A Method of Case Structure Analysis for Japanese Sentences based on Examples in Case Frame Dictionary. *IEICE Transactions on Information and Systems*, Vol. E77–D, No. 2, pp. 227–239, 1994.

(12) 村田真樹, 長尾眞. 表層表現と用例を用いた照応省略解析手法. 言語理解とコミュニケーション研究会 NLC, 1998.

(13) 山下達雄, 松本裕治. 品詞タグ付きコーパスを直接利用した形態素解析. 言語理解とコミュニケーション研究会 NLC, 1998.

(14) Jiri Stetina and Makoto Nagao. PP Attachment Ambiguity Resolution through Supervised Learning. *Journal of Natural Language Processing*, Vol. 5, No. 1, 1998.

(15) 矢田恭儀. 用例とシソーラスからの決定木学習による名詞句「a の b」の意味解析. 京都大学工学部修士論文, 1997.

(16) 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会 第 3 回年次大会, pp. 115–118, 1997.

(17) 黒橋禎夫. 日本語構文解析システム KNP 使用説明書 version 2.0b4. 京都大学大学院情報学研究科, 1997.

(18) 黒橋禎夫. 日本語構文解析システム KNP 使用説明書 version 2.0b6. 京都大学大学院情報学研究科, 1998.